Re-Identification of Zebrafish using Metric Learning

Haurum, Joakim Bruslund; Karpova, Anastasija; Pedersen, Malte; Bengtson, Stefan Hein; Moeslund, Thomas B.

Published in:
Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW 2020

DOI (link to publication from Publisher):
10.1109/WACVW50321.2020.9096922

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Re-Identification of Zebrafish using Metric Learning

Joakim Bruslund Haurum* Anastasija Karpova*
Malte Pedersen Stefan Hein Bengtson Thomas B. Moeslund
Visual Analysis of People (VAP) Laboratory, Aalborg University, Denmark
{joha,mape,shbe,tbm}@create.aau.dk, akkarpova@gmail.com

Abstract

Zebrafish are widely used for drug development and behavioral pattern studies. The currently employed zebrafish re-identification methods rely solely on top-view and grayscale images which require a significant amount of annotated data in order to perform well. In this paper, for the first time, we perform zebrafish re-identification using RGB images recorded from a side-view perspective, while keeping the amount of data annotation to a minimum. Inspired by the person re-identification field, two feature descriptors are tested, each encoding both color and texture information, and five metric and subspace learning methods. The contribution of the color and texture components of the feature descriptors were also investigated. We present and evaluate on a novel publicly available dataset of six zebrafish, recorded in a laboratory setup. The results show that a mean average precision of 99% can be achieved by using just 15 annotated samples per fish. This approach shows a clear potential for incorporating the side-view information in the field of zebrafish tracking, as well as a clear argument for utilizing the color information of the zebrafish.

1. Introduction

The zebrafish (*Danio rerio*) has for many years been used as a vertebrate model organism by biologists. This has been due to a major effort in screening the zebrafish genomes [11, 16] and their transparent body during early development, allowing non-obtrusive observation of the subjects [17]. Due to these properties zebrafish have been used to study the effect of drugs [14], complex brain disorders [20], and more. The zebrafish is also a highly social animal, reflected in the shoaling behaviour observed under various conditions [32], which affects the proximity between the fish, and the direction and speed of the individual fish. In order to properly analyze these results, it is necessary to track each unique fish over time. This is, however, an incredible difficult task as the zebrafish behave in an impulsive and unpredictable way when stressed or anxious [10]. Therefore, zebrafish have traditionally been tracked manually by the researchers [22, 30, 33].

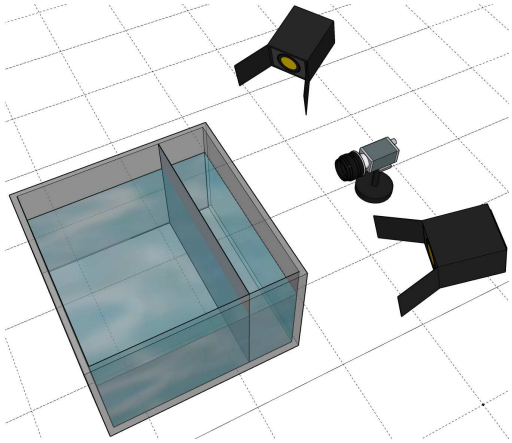


Figure 1: Illustration of the dataset capture setup. The equipment size and distances are not to scale.

In recent years there has, however, been a growing effort in creating automated zebrafish tracking systems. Systems such as the idTracker [38, 45] have been widely used in research, and several commercial systems are being sold [28, 34, 49, 50]. These tracking systems are constrained to only observe the fish in approximated 2D planes. Therefore, they only allow the fish to swim in very shallow water, which severely limits the movement of the fish. This is a major problem, as MacRí *et al*. [29] found that concluding on zebrafish behaviors determined from 2D data is not representative of the actual zebrafish behavior. Comparatively, utilizing 3D data in order to conclude on zebrafish behavior provides much more reliable results. Tracking zebrafish in 3D is, however, a much more difficult task due to an increase in occlusions, variety of body poses, and more. This leads to less stable tracking and thereby more track-

*Equal contribution

lets, which subsequently needs to be combined. One way to combine these tracklets would be through re-identifying the zebrafish in the tracklets, assuming the tracklets do not contain identity swaps. This approach has been applied from a top-view perspective [9, 38, 45, 54, 59], but never from a side-view perspective, due to the perceived increase in occlusions leading to a more difficult tracking problem. Furthermore, the color information has not been utilized by any of the current methods, representing an uncharted area within the field. Lastly, there is a distinct lack in public ground truth annotated data within the field, making it exceedingly hard to compare methods. Therefore, our contributions are the following:

- Demonstrating that zebrafish can reliably be re-identified in a side-view perspective.
- A publicly available side-view dataset of zebrafish, with bounding box annotations, recorded in color, and with temporally consistent IDs[1].
- Open-source python implementations of the applied re-identification methods and feature descriptors[2].

2. Related Work

When attempting to track multiple objects over long durations in non-trivial circumstances, it is often necessary to handle occlusions by re-identifying and re-assigning the involved objects. This is not a trivial task as illustrated by the exponentially increasing number of re-identification papers accepted at major computer vision conferences over the last decade [63]. Over the years the field of person re-identification has heavily utilized the fields of feature engineering, and metric and subspace learning. Through metric and subspace learning it is possible to learn a set of transformations on the feature space, with the goal of minimizing the intra-class distances while maximizing the inter-class distances, by modelling the *Mahalanobis matrix* [7].

Weinberger and Saul [55] proposed an iterative approach where a local perimeter is enforced for each sample wherein only samples of the same class may be contained. Köstinger *et al*. [21] proposed a simple approach based on the pairwise difference for similar and dissimilar samples, called Keep It Simple and Straightforward Metric (KISSME), from which the Mahalanobis matrix could analytically be determined. This approach was subsequently expanded on by Yang *et al*. [61] and Liao *et al*. [23], who incorporated pairwise commonness and a subspace learning step, respectively. On the other hand, Zhang *et al*. [62] applied a subspace learning approach called Discriminative Null Space (DNS) to learn a heavily reduced feature subspace wherein the different classes are well separated, by using non-linear transformation through a kernelization approach.

Convolutional Neural Networks (CNNs) have had a large impact on the field, where they have been used to learn feature extractors [56, 58] and end-to-end metric learning [1, 5, 66]. In recent years the field has pivoted towards developing algorithms which adapts to never before seen areas. This has been done by utilizing one-shot learning [4], Generative Adversarial Networks [25, 64], domain adaptation [48, 65], and weakly-supervised and unsupervised methodologies [31, 60, 65].

While person re-identification has increased in popularity the field of animal re-identification has not followed an as rapidly increasing attention. In a recent review by Schneider *et al*. [47] it is clear that animal re-identification has a long history, dating back to 1990, exploring different feature and machine learning based approaches on a large variety of species. In recent years deep learning has had a large impact on the field, leading to increased performance and attention. Similarly, the introduction of publicly available datasets and challenges such as the Caltech Camera Traps [6] and the Humpback Whale Identification Challenge [19] have led to an increased interest in the field. The produced technology has been widely used for species conservation and censusing, by incorporating citizen science to help gather data [37], and developing initiatives such as the Wildbook project [8] which has eased and improved the entire conservation process immensely.

However, image-based re-identification has only been seldom used in the field of zebrafish tracking. Several tracking systems have been proposed, with the majority of the systems focusing on 2D tracking of zebrafish in shallow water [38, 41, 42, 45, 51, 54, 59]. Only a relatively small amount of work has been conducted in the attempt to create reliable 3D tracking systems [2, 9, 27, 39, 40, 52, 53, 57]. Within all of these systems only five attempt to explicitly model the appearance of the fish, and all from the top-view camera. Cheng *et al*. [9, 59] applied an iterative unsupervised method to train a CNN to re-identify the fish based on head patches. Similarly, Wang *et al*. [54] applied a CNN to re-identify zebrafish heads in a supervised manner. Pérez-Escudero *et al*. [38] proposed an identification method based on intensity and contrast maps in a system called idTracker. Romero-Ferrero *et al*. [45] proposed a new version of idTracker, where the identification step was performed using a small classification CNN. In all cases the utilized data has been recorded in grayscale from a top-view camera, as to limit the number of occlusions. Everyone tracks the zebrafish in 2D using a single top-view camera, except Cheng *et al*. who tracks the fish in 3D using a triple camera setup. The seemingly unique and contrasting stripes of the zebrafish have therefore never been utilized. Sim-

[1] https://www.kaggle.com/aalborguniversity/aau-zebrafish-reid

[2] https://www.bitbucket.org/aauvap/zebrafish-re-identification

ilarly, there is a distinct lack in publicly available ground truth annotated datasets, which makes it exceedingly hard to compare methodologies, and we believe this may have slowed down progress in the field when comparing to the otherwise rapid progress within person re-identification.

3. Dataset

The dataset was recorded strictly from a side-view perspective in a laboratory environment. A sketch of the laboratory dataset collection setup is shown in Figure 1. The recorded dataset is intended specifically for the re-identification task.

3.1. Experimental Setup

A $32 \times 32 \times 32$ cm clear glass tank was used, with a water depth of 10.5 cm, and a clear acrylic plate inserted 3.5 cm from the front. The divider plate was inserted in order to limit the depth-wise movement of the zebrafish, forcing the fish to swim at approximately the same distance from the camera. The divider plate also forces the fish to swim approximately perpendicular to the camera, by allowing the fish to turn but not to swim towards or away from the camera. Two Kino Diva-Lite 401-230 studio lamps with fluorescent tube lights and a refresh rate of 40 kHz were used, in order to avoid flicker when recording with a high shutter speed and ensure a smooth lighting of the fish tank. The lights were placed 90 cm from the tank at an approximate 45-degree angle. The lighting was diffused by placing the fish tank in a photography tent, limiting the amount of oversaturated highlights on the fish scales.

The videos were recorded in RGB using an IDS UI-3070CP Rev. 2 camera and a KOWA LM16HC lens, with a resolution of 2056×1542, a variable frame rate, exposure time of 9.175 ms, and a manually set color balance. The camera setup was placed 70 cm from the front of the fish tank, perpendicular to the water level in the tank. The KOWA lens was chosen in order to obtain a narrow field-of-view, limiting the visibility of the sides of the tank and thereby eliminating any reflections on the side of the tank.

3.2. Dataset Construction

A total of six unique zebrafish were recorded, each shown in Figure 2. Due to the limited space of the tank, two videos with three fish at a time were recorded, and combined into a single dataset of 2224 images. Each recording was manually annotated with bounding boxes and unique consistent ids throughout the video, using the AAU VAP Bounding Box Annotator software [3]. For each bounding box it was denoted whether the fish was swimming to the right or left, turning/swimming at an angle, or part of an occlusion.

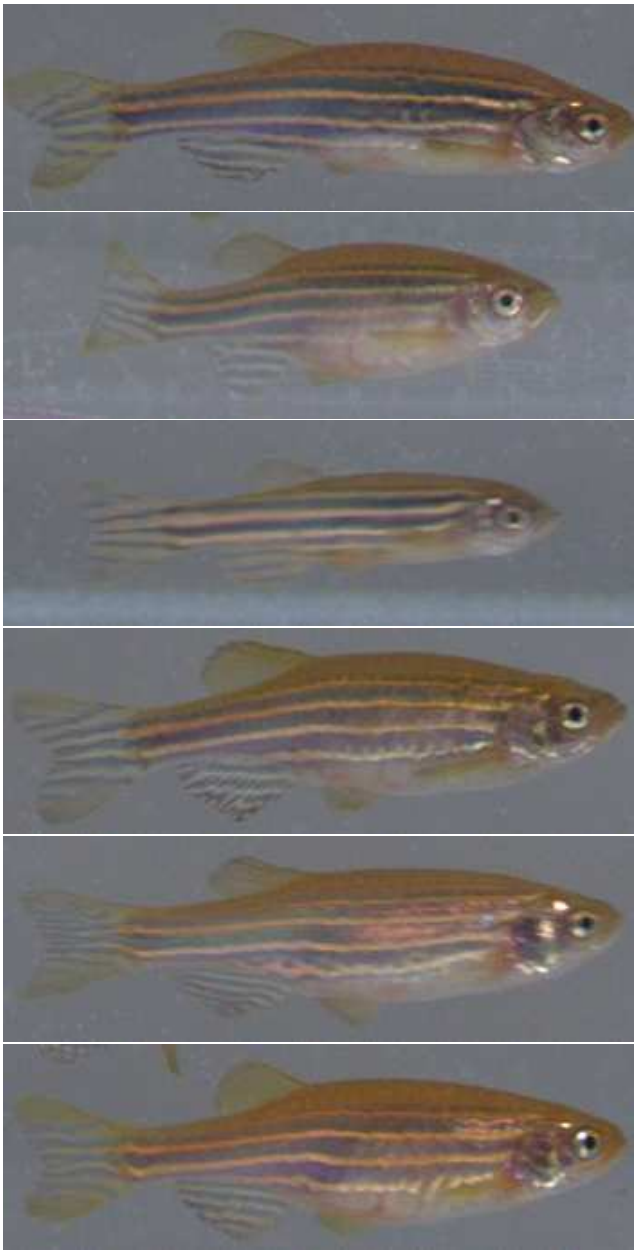
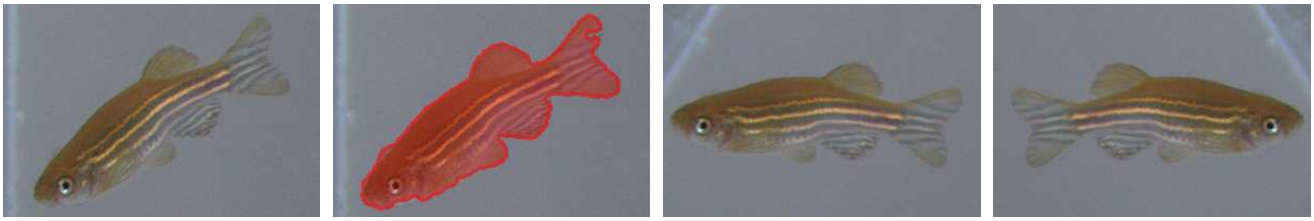


Figure 2: Still images of each of the 6 different zebrafish in the recorded dataset.

4. Methodology

In order to determine the feasibility of re-identifying zebrafish from a side-view perspective, we investigate five analytic metric and subspace-learning methods:

- Keep It Simple and Straightforward Metric (KISSME) [21]
- Improved KISSME (iKISSME) [61]
- Large Scale Similarity Learning (LSSL) [61]



(a) Initial fish patch (b) Segmented fish (c) Rotated onto x-axis (d) Flipped fish patch

Figure 3: An example of the applied pre-processing steps. First the fish BLOB is estimated, where after the BLOB is rotated onto the x-axis, and lastly flipped if the fish head is in the left half of the image patch.

- Cross-view Quadratic Discriminant Analysis (XQDA) [23]
- Kernelized Discriminative Null Space (DNS) [62]

As these methods require a pre-computed feature descriptor, we also investigate two feature descriptors:

- Ensemble of Localized Features (ELF) [13, 56]
- Local Maximal Occurrence (LOMO) [23]

4.1. Pre-processing

It is assumed that the bounding box and direction of the head of the fish have been determined a priori. As the fish may swim in a diagonal fashion, the axis-aligned bounding boxes can contain a lot of empty space, which provides no relevant information. Furthermore, it is assumed that each side of the fish is non-significantly different. Therefore, we want to rotate the bounding boxes so all extracted fish are approximately parallel to the x-axis. Subsequently, the images are flipped so that the head of the fish points in the same direction in all the image patches. These steps are shown in Figure 3 and performed as follows:

Based on the detected bounding box the fish is segmented through a simply process of median background subtraction, thresholding, and morphological opening and closing. In case several objects are present within the bounding box only the BLOB with the largest area is kept. The angle of the zebrafish to the x-axis is determined by computing the eigenvectors of the segmentation mask, and rotating the eigenvector with the largest eigenvalue onto the x-axis. The new bounding box is then determined, and the image patch flipped if the fish head is pointing the wrong way. Lastly, all extracted fish patches are resized into a single consistent size.

4.2. Feature Descriptors

Two different feature descriptors from the person re-identification field were utilized: ELF and LOMO. Both descriptors consist of a color and texture component, and constructs the descriptor based on a horizontal stripe analysis approach.

4.2.1 ELF

The ELF feature descriptor is a simple descriptor constructed using several different color spaces and texture response filters. The color component consists of RGB, YCbCr, and HSV color spaces, however, as the Y and V channels of the YCbCr and HSV color spaces are identical only the Y channel is used, resulting in eight color channels. The texture component consists of the filter responses on the Y channel when applying the Schmid [46] and Gabor [12] filter banks, as well as calculating the standard Local Binary Pattern [35, 36], resulting in 22 texture channels.

The final feature representation is created by splitting the channels into six horizontal stripes of equal size, and for each stripe represent each channel with an ℓ_1 normalized 16-bin histogram. Per stripe all histograms are concatenated, and the final six stripe histogram vectors are concatenated into a single feature vector.

4.2.2 LOMO

The LOMO feature descriptor was developed in order to provide a scale and pose invariant representation, as different poses had typically caused problems for person re-identification tasks. This is achieved by utilizing an overlapping patch based approach, where for each patch a joint-HSV histogram and a Scale Invariant Local Ternary Pattern (SILTP) [24] histogram is calculated. In order to make the feature descriptor scale invariant, a three-scaled pyramid representation is constructed by applying a 2×2 mean filter per step. In order to make the feature descriptor pose invariant, each row of patches is analyzed and the "local maximal occurrence" is determined by only selecting the largest bin value across all patch histograms in the row.

All row histograms are concatenated, and subsequently concatenated across the three scale steps. Finally, large values in the joint-HSV and SILTP histogram vectors are suppressed using the log operator, the vectors are ℓ_2 normalized and concatenated into a single feature vector. In the original implementation the multi-scale Retinex transformation [18] was applied as a pre-processing step in order to match

lighting and colors from different cameras. As we utilize a single camera setup this transformation is not applied.

4.3. Metric Learning

A central problem for classification, ranking, and re-identification problems is to determine whether two inputs (*e.g*. two faces or two pedestrian images) are similar or not. This can be done through distance metric or similarity measure between two feature representations. A distance metric is a pairwise real-valued function with two d-dimensional vectors, $\mathbf{x}$ and $\mathbf{y}$, as input, that obeys the following conditions [7]:

1. Non-negative: $f(\mathbf{x}, \mathbf{y}) \geq 0$
2. Symmetric: $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$
3. Triangle inequality: $f(\mathbf{x}, \mathbf{z}) \leq f(\mathbf{x}, \mathbf{y}) + f(\mathbf{y}, \mathbf{z})$
4. Identity of indiscernibles: $f(\mathbf{x}, \mathbf{y}) = 0$ iff. $\mathbf{x} = \mathbf{y}$

The advantage of using metric learning is a joint optimization of feature representation and the deciding metric. One of the fundamental learned metrics is the Mahalanobis metric, [7], which can be used to determine whether $\mathbf{x}$ and $\mathbf{y}$ are from the same distribution, parameterized by the covariance matrix $\mathbf{\Sigma}$, see Equation 1. It should be noted that $\mathbf{M}$ is often used as shorthand for $\mathbf{\Sigma}^{-1}$, and should be Positive Semi-Definite (PSD) in order for the Mahalanobis metric, d_{M}, to be a pseudo-metric.

$$d_{\mathrm{M}}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^{\mathbf{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})} \tag{1}$$

The KISSME algorithm calculates $\mathbf{M}$ by modelling the pairwise difference between similar and dissimilar points as two separate zero-centered Gaussian distributions. $\mathbf{M}$ is then simply determined as the difference of the inverse of the computed covariance matrices.

Yang *et al*. [61] proposed the LSSL method, where a similarity measure is computed by utilizing both the pairwise difference and commonness between similar and dissimilar feature representations. Based on these properties two matrices which parametrizes the similarity and dissimilarity measures are constructed. Yang *et al*. also find that the $\mathbf{M}$ used in the KISSME approach can be determined utilizing only the pairwise difference between similar points, leading to the iKISSME algorithm.

Liao *et al*. [23] expanded on the KISSME algorithm by learning a transformation which projects the feature points into a smaller subspace. This transformation is found by solving the generalized eigenvalue decomposition problem given the covariance matrices for the pairwise differences for the similar and dissimilar points.

Zhang *et al*. [62] proposed a subspace learning algorithm called DNS for the small sample size problem, which approaches the case where only k d-dimensional samples are available per class and that $k \ll d$. The goal of the DNS algorithm is to find a subspace wherein the intra-class scatter matrix is zero, while the inter-class scatter matrix is non-zero. This is achieved by using the Null Foley-Sammon Transformation [15], which finds a $c - 1$ dimensional subspace, where c is the number of classes considered. Within this heavily reduced subspace the distance between samples are simply calculated using the common ℓ_2 distance. Zhang *et al*. further developed a kernelized version of the algorithm, which allows learning non-linear transformations of the feature space by applying different kernels such as the Radial Basis Function (RBF) kernel.

5. Experimental Results

The feature descriptors and metric learning methods described in Section 4 are evaluated. Additionally, a baseline performance is established by simply measuring the distance between feature vectors of the samples utilizing ℓ_1, ℓ_2, and the cosine distances. The metric learning methods were tested in two scenarios: with and without forcing the Mahalanobis matrix, $\mathbf{M}$, onto the PSD cone. The DNS method was tested with both the linear kernel and the RBF kernel.

Similarly, in order to determine the effect of the color and texture components of the feature descriptors, three variations of the descriptors are investigated: the full descriptor (LOMO / ELF), only the color components (LOMO-HSV / ELF-COLOR), and only the texture component (LOMO-SILTP / ELF-TEXTURE). For all feature descriptors the input data was rotated and flipped, so all fish bounding boxes were parallel to the horizontal axis, with the head pointing to the right.

5.1. Dataset Split and Evaluation Metrics

The methods were evaluated on the recorded dataset described in Section 3, where a perfect fish detector is assumed. The feature descriptions are therefore extracted from the ground truth annotations, which have been resized to the median bounding box size, 330×99. Therefore, the extracted ELF feature descriptor is represented in a 2880 dimensional space, whereas the LOMO descriptor is represented in a 19546 dimensional space. The color information is encoded in a 768 and 14848 dimensional subspace for ELF and LOMO, respectively, whereas the texture information is encoded in a 2112 and 4698 dimensional subspaces for ELF and LOMO, respectively. Cases where the fish is turning or occluding each other are excluded. The evaluation is cross-validated across ten random splits. The splits were constructed so that each unique fish has an equal number of samples. The fish with the least number of valid annotations, given the previously imposed restrictions, determines the size of the data splits. Therefore, each split consists of 583 randomly selected bounding boxes per fish. Per split 100 samples are selected per fish as the training

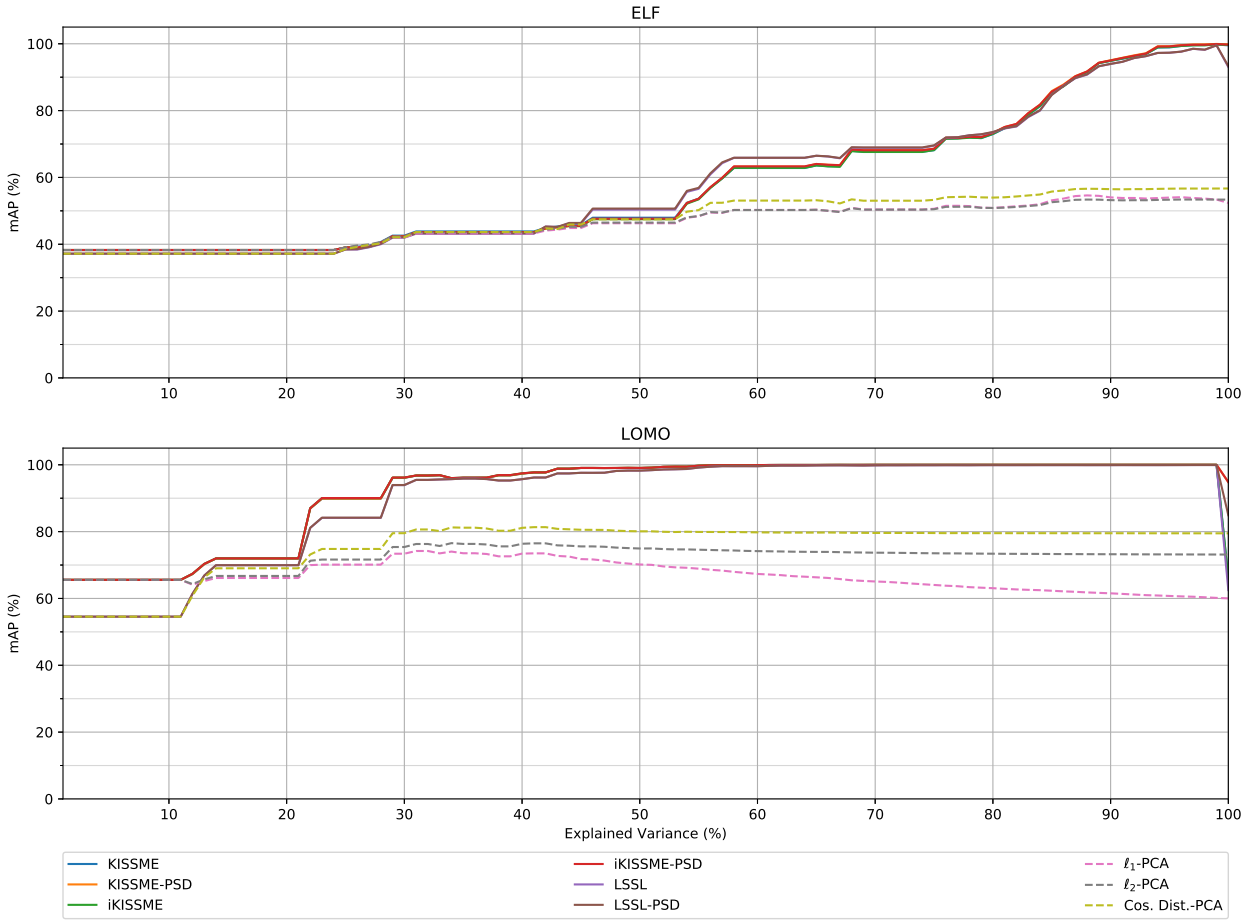


Figure 4: mAP results for the PCA hyperparameter search across the relevant re-identification methods and baselines for the two full feature descriptors, compared with the amount of explained variance kept during the PCA process. All tests are performed with 100 training samples per fish, using three fish and 10-fold cross-validation. The baseline methods are plotted with dashed lines.

set, leaving the remaining 483 samples as the testing set. From the testing set a single sample per fish is selected as the *probe* sample, leaving 482 *gallery* samples. The performance of the methods is measured using the mean Average Precision (mAP) metric. The classic Cumulative Matching Criteria (CMC) metric is not utilized, as it does not reflect the accuracy of the tested methods given several gallery samples per id.

5.2. Hyperparameter Selection

For some of the investigated methods a set of hyperparameters needs to be considered. In order to simplify the tests, the original parameters are used for each method, unless otherwise stated.

The KISSME, iKISSME, and LSSL methods all include a matrix inversion step when computing $\mathbf{M}$. In order to make this process feasible given the large feature descriptors used, it is necessary to apply some kind of dimensionality reduction. As in the original papers we apply Principal Component Analysis (PCA). In order to determine the extent of the dimensionality reduction, the tradeoff between accounted variance and mAP performance is considered. This is evalauted over a subset of the full dataset, where only three fish ids are used in ten different splits with 100 training features, and only comparing the full ELF and LOMO descriptors. The effect on the mAP while increasing the amount of explained variance from 1% to 100% in steps of 1% is measured. This is shown in Figure 4, where we also plot the baseline methods performance using the PCA reduced feature descriptors. We find that only 60% and 95% of the explained variance should be included for LOMO and ELF, respectively, in order to obtain the peak mAP performance.

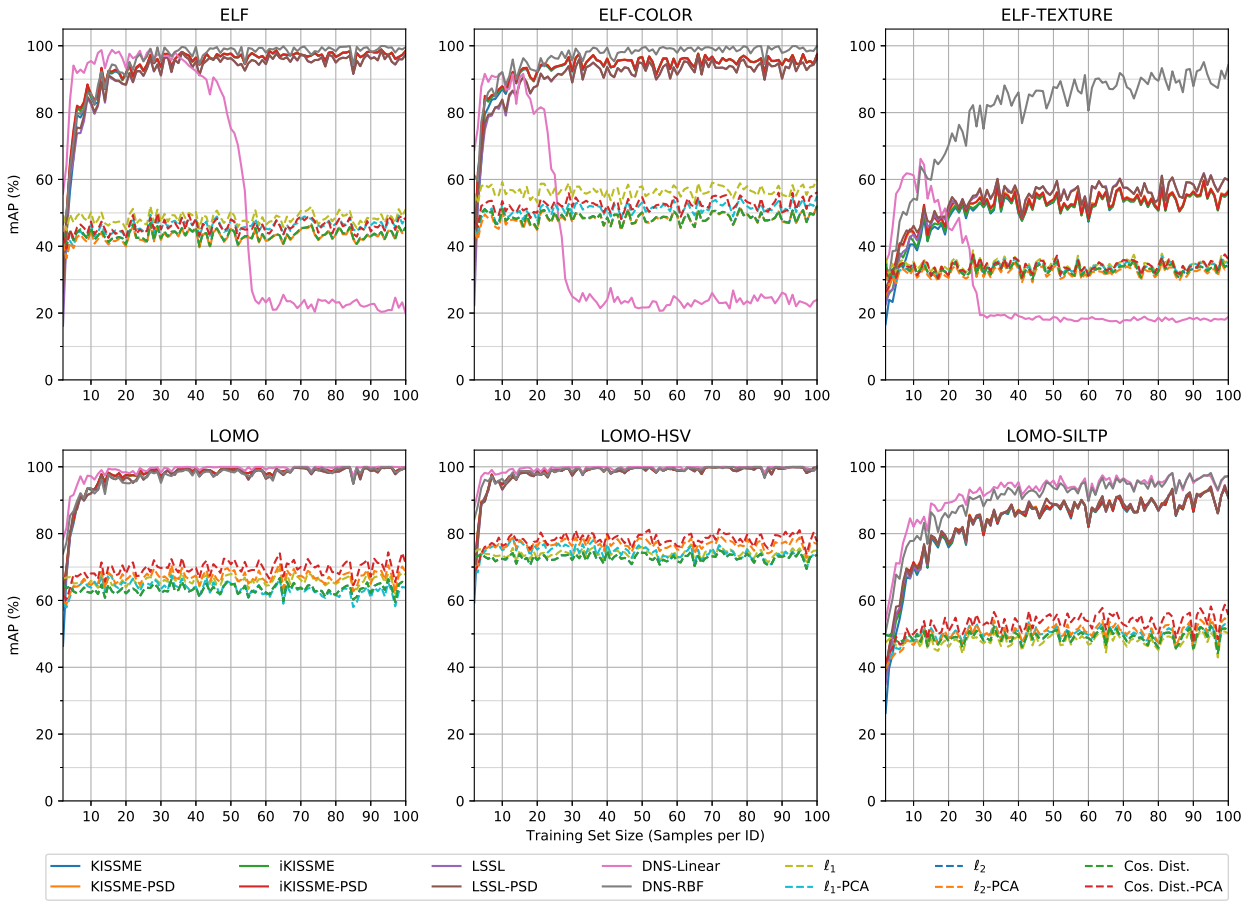


Figure 5: mAP result for all tested methods and feature descriptors, over varying training set sizes, denoted as samples per fish. The average mAP over 10-fold cross-validation is reported. Please note that the XQDA method is not included, as the method did not manage to calculate a meaningful subspace transformation. The baseline methods are plotted with dashed lines, and tested using both the full feature descriptors and the dimensionality reduced descriptors.

5.3. Effect of Training Set Size

As annotated data is often a scarce and expensive resource, it is of great interest to investigate the effect of the training set size on the performance of the methods. Therefore, we investigate the effect of the training set size by evaluating all methods and feature descriptor combinations. We conduct this test by evaluating the training set size from the bare minimum of two to the maximum possible of 100. For each increment an additional training feature per fish is added to the previous set of features. The results are shown in Figure 5, where each method is represented as the mean mAP across the ten splits for each training set size. Please note that the XQDA method is not included as the method was never capable of finding a usable subspace from the training data. The baselines are evaluated using both the full and dimensionality reduced feature descriptors.

6. Discussion

From the results it is clear that all of the tested re-identification methods are capable of achieving near perfect results with just a small training set, as an mAP of 99% is achieved using just 15 samples per fish. When comparing with the baseline metrics it is also apparent that the feature space transformations encoded in the Mahalanobis matrix and learned subspace (for DNS) has a large effect, leading to a 20-40 percentage points increase in mAP.

When examining the results, the DNS method consistently performs better in terms of achieving a higher mAP with less data. This is true for both the linear and RBF versions of the algorithm. However, the linear version of the DNS algorithm diverges when the training set size is increased with the ELF feature descriptor. This may be related to the feature space being too non-linear, as the ELF feature representation is encoded in a significantly smaller

feature space when compared to the LOMO feature representation.

When determining the effects of the different feature representation components, it is obvious that the color components have a much larger effect than the texture component. This is somewhat surprising as one's initial intuition would expect the well-defined striped textures of the zebrafish to be a major feature of the fish, as proven by the majority of the work in the field being conducted on grayscale data. For the ELF feature descriptor only the RBF kernelized DNS algorithm manages an mAP over 60%, while the remaining methods perform significantly worse, though still better than the baseline methods. The effect is not as pronounced for the texture component of LOMO, though the DNS algorithm still consistently outperforms all of the other methods. The superior contribution of the color component is similarly clear when looking at the baselines. The baselines evaluated on just the color components of the feature descriptors consistently outperforms not only the texture components, but also the full feature descriptors, achieving an mAP of 70-80% for the LOMO-HSV descriptor, and 50-60% for the ELF-COLOR descriptor. While this effect can also be seen to a small degree for some of the learned methods, it is much more pronounced for the baseline methods. Furthermore, given a training set size of 20-30 samples per fish it appears that the difference in performance between using the full feature descriptor and only the color component becomes negligible. Based on these observations a set of conclusions can be made.

If the best performance is needed, the results indicate that utilizing the LOMO-HSV feature descriptor with the DNS algorithm with either of the two kernels gives the best performance with the lowest amount of training data. However, the LOMO feature descriptor is more computationally heavy to compute than the ELF descriptor, due to its complex structure and detailed patch construction. The ELF descriptor would therefore be favourable in time critical systems.

Similarly, since any kernelized method explicitly needs the training data to be stored in order to calculate the test data kernel matrix, the DNS algorithm may not be the best choice for systems with low storage capabilities. In these cases, an approximate performance can be achieved by using any of the metric learning based methods when using LOMO, given a training set size of 20 or more samples per fish. However, if ELF is used a small performance drop in mAP is to be expected.

As mentioned earlier the required training set size to reach an mAP of 99% is only 15 samples per fish. This is on such a small scale that it would be reasonable to ask experts, *i.e*. biologists, to manually annotate a small set of images per fish. Similarly, it would be reasonable to expect that a tracker could produce a tracklet consisting of at least 15 frames, from which the underlying model can be calculated. Comparatively, if a CNN based approach was used the required amount of annotated data would be much higher, and may not necessarily generalize to the different test environments. While the current methods do not generalize to new identities, they are quick to train, making the lack of generalization non-significant. However, there is a set of unanswered questions that should be considered.

Currently, a detector is assumed which provides perfect annotations of fish that are swimming perpendicular to the camera. Furthermore, the fish are currently forced to swim within a small plane of water, limiting the light scattering effect through the water. It is therefore of great interest to study how these methods work when provided with feature descriptors extracted from imperfect bounding box detections from modern detection networks such as YOLO [43], Faster-RCNN [44], or SSD [26], and how well the methods perform when color distortions and reduction of detail is introduced by letting the fish swim in three dimensions.

7. Conclusion

In this work we have addressed how methods from the person re-identification field can be used in order to reliably re-identify zebrafish. A novel RGB dataset with six zebrafish was recorded in a lab environment using a single side-view camera setup, where the fish were constrained to swim within a 3.5 cm plane at the front of the tank. Based on the recorded dataset two feature descriptors and five analytic metric and subspace learning methods are compared under varying training set sizes. The test is conducted with 10-fold cross-validation, and the results indicate that it is possible to achieve a mean Average Precision (mAP) of 99% with just 15 training samples per class that. Furthermore, in-depth analysis of the feature descriptors shows that the main contributor to the recognition performance is the color component, and not the texture component. The results further indicate that by just using the color component, a higher mAP value can be achieved using less training data when compared with the full feature descriptors. This is in stark contrast to the traditional approach within zebrafish tracking where color information is discarded during data acquisition. These results clearly indicate that there is valuable information to be utilized from the side-view perspective which is currently rarely used, and that the color information should not simply be discarded, unlike the current practice. In the future it would be of great interest to investigate how the applied re-identification methods work in a full 3D tracking system, where the assumptions of perfect bounding box detections and lack of color distortions from the water are not met. Similarly, it would be interesting to investigate whether the investigated re-identification methods are capable of re-identifying the same fish across data recorded at different days.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[2] G. Audira, B. Sampurna, S. Juniardi, S.-T. Liang, Y.-H. Lai, and C.-D. Hsiao. A simple setup to perform 3D locomotion tracking in zebrafish by using a single camera. *Inventions*, 3(1):11, Feb. 2018.

[3] C. H. Bahnsen, A. Møgelmose, and T. B. Moeslund. The aau multimodal annotation toolboxes: Annotating objects in images and videos. *arXiv preprint arXiv:1809.03171*, 2018.

[4] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] S. Bak and P. Carr. Deep deformable patch metric learning for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2690–2702, Oct 2018.

[6] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 472–489, Cham, 2018. Springer International Publishing.

[7] A. Bellet, A. Habrard, and M. Sebban. *Metric Learning*. Morgan & Claypool, 2015.

[8] T. Y. Berger-Wolf, D. I. Rubenstein, C. V. Stewart, J. A. Holmberg, J. Parham, S. Menon, J. P. Crall, J. V. Oast, E. Kiciman, and L. Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *CoRR*, abs/1710.08880, 2017.

[9] X. E. Cheng, S. S. Du, H. Y. Li, J. F. Hu, and M. L. Chen. Obtaining three-dimensional trajectory of multiple fish in water tank via video tracking. *Multimedia Tools and Applications*, 77(18):24499–24519, Feb. 2018.

[10] R. J. Egan, C. L. Bergner, P. C. Hart, J. M. Cachat, P. R. Canavello, M. F. Elegante, S. I. Elkhayat, B. K. Bartels, A. K. Tien, D. H. Tien, S. Mohnot, E. Beeson, E. Glasgow, H. Amri, Z. Zukowska, and A. V. Kalueff. Understanding behavioral and physiological phenotypes of stress and anxiety in zebrafish. *Behavioural Brain Research*, 205(1):38 – 44, 2009.

[11] J. S. Eisen. Zebrafish make a big splash. *Cell*, 87(6):969–977, Dec. 1996.

[12] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Biological Cybernetics*, 61(2):103–113, Jun 1989.

[13] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision – ECCV 2008*, pages 262–275, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[14] J. Green, C. Collins, E. J. Kyzar, M. Pham, A. Roth, S. Gaikwad, J. Cachat, A. M. Stewart, S. Landsman, F. Grieco, R. Tegelenbosch, L. P. Noldus, and A. V. Kalueff. Automated high-throughput neurophenotyping of zebrafish social behavior. *Journal of Neuroscience Methods*, 210(2):266 – 271, 2012.

[15] Y.-F. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue. Null foley–sammon transform. *Pattern Recognition*, 39(11):2248 – 2251, 2006.

[16] P. Haffter, M. Granato, M. Brand, M. Mullins, M. Hammerschmidt, D. Kane, J. Odenthal, F. van Eeden, Y. Jiang, C. Heisenberg, R. Kelsh, M. Furutani-Seiki, E. Vogelsang, D. Beuchle, U. Schach, C. Fabian, and C. Nusslein-Volhard. The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development*, 123(1):1–36, 1996.

[17] A. J. Hill, H. Teraoka, W. Heideman, and R. E. Peterson. Zebrafish as a Model Vertebrate for Investigating Chemical Toxicity. *Toxicological Sciences*, 86(1):6–19, 02 2005.

[18] D. J. Jobson, Z. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7):965–976, July 1997.

[19] Kaggle. Humpback whale identification challenge. `https://www.kaggle.com/c/whale-categorization-playground`, 2018. Accessed: 2019-12-09.

[20] A. V. Kalueff, A. M. Stewart, and R. Gerlai. Zebrafish as an emerging model for studying complex brain disorders. *Trends in Pharmacological Sciences*, 35(2):63 – 75, 2014.

[21] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large Scale Metric Learning from Equivalence Constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[22] L. Li and J. E. Dowling. A dominant form of inherited retinal degeneration caused by a non-photoreceptor cell-specific mutation. *Proceedings of the National Academy of Sciences*, 94(21):11645–11650, Oct. 1997.

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1301–1306, June 2010.

[25] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang. Adaptive transfer network for cross-domain person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing.

[27] X. Liu, Y. Yue, M. Shi, and Z.-M. Qian. 3-D video tracking of multiple fish in a water tank. *IEEE Access*, 7:145049–145059, 2019.

[28] Loligo Systems. LoliTrack v.4. `https://www.loligosystems.com/lolitrack-v-4`.

[29] S. Macrì, D. Neri, T. Ruberto, V. Mwaffo, S. Butail, and M. Porfiri. Three-dimensional scoring of zebrafish behavior

unveils biological phenomena hidden by two-dimensional analyses. *Scientific Reports*, 7(1), May 2017.

[30] M. B. McElligott and D. M. O'Malley. Prey tracking by larval zebrafish: Axial kinematics and visual control. *Brain, Behavior and Evolution*, 66(3):177–196, 2005.

[31] J. Meng, S. Wu, and W.-S. Zheng. Weakly supervised person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[32] N. Miller and R. Gerlai. From schooling to shoaling: Patterns of collective motion in zebrafish (danio rerio). *PLOS ONE*, 7(11):1–6, 11 2012.

[33] U. K. Muller. Swimming of larval zebrafish: ontogeny of body waves and implications for locomotory development. *Journal of Experimental Biology*, 207(5):853–868, Feb. 2004.

[34] L. P. J. J. Noldus, A. J. Spink, and R. A. J. Tegelenbosch. EthoVision: A versatile video tracking system for automation of behavioral experiments. *Behavior Research Methods, Instruments, & Computers*, 33(3):398–414, Aug. 2001.

[35] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585 vol.1, Oct 1994.

[36] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, July 2002.

[37] J. Parham, J. Crall, C. Stewart, T. Berger-Wolf, and D. Rubenstein. Animal population censusing at scale with citizen science and photographic identification. In *AAAI Spring Symposium Series*, 2017.

[38] A. Pérez-Escudero, J. Vicente-Page, R. C. Hinz, S. Arganda, and G. G. de Polavieja. idTracker: tracking individuals in a group by automatic identification of unmarked animals. *Nature Methods*, 11(7):743–748, June 2014.

[39] Z. Qian, M. Shi, M. Wang, and T. Cun. Skeleton-based 3D tracking of multiple fish from two orthogonal views. In *Communications in Computer and Information Science*, pages 25–36. Springer Singapore, 2017.

[40] Z.-M. Qian and Y. Q. Chen. Feature point based 3D tracking of multiple fish from multi-view images. *PLOS ONE*, 12(6):1–18, 2017.

[41] Z.-M. Qian, X. E. Cheng, and Y. Q. Chen. Automatically detect and track multiple fish swimming in shallow water with frequent occlusion. *PLOS ONE*, 9(9):1–12, 2014.

[42] Z.-M. Qian, S. H. Wang, X. E. Cheng, and Y. Q. Chen. An effective and robust method for tracking multiple fish in video image based on fish head detection. *BMC Bioinformatics*, 17(1):251, June 2016.

[43] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[44] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[45] F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. H. Heras, and G. G. de Polavieja. idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nature Methods*, 16(2):179–182, Jan. 2019.

[46] C. Schmid. Constructing models for content-based image retrieval. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II, Dec 2001.

[47] S. Schneider, G. W. Taylor, S. Linquist, and S. C. Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019.

[48] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[49] TSE Systems. VideoMot2 - versatile video tracking system. `https://www.tse-systems.com/product-details/videomot`.

[50] ViewPoint. ZebraLab. `http://www.viewpoint.fr/en/p/software/zebralab`.

[51] S. H. Wang, X. E. Cheng, and Y. Q. Chen. Tracking undulatory body motion of multiple fish based on midline dynamics modeling. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2016.

[52] S. H. Wang, X. Liu, J. Zhao, Y. Liu, and Y. Q. Chen. 3D tracking swimming fish school using a master view tracking first strategy. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Dec. 2016.

[53] S. H. Wang, J. Zhao, X. Liu, Z. Qian, Y. Liu, and Y. Q. Chen. 3D tracking swimming fish school with learned kinematic model using LSTM network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1068–1072, March 2017.

[54] S. H. Wang, J. W. Zhao, and Y. Q. Chen. Robust tracking of fish schools using CNN for head identification. *Multimedia Tools and Applications*, 76(22):23679–23697, Nov. 2016.

[55] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009.

[56] S. Wu, Y. Chen, X. Li, A. Wu, J. You, and W. Zheng. An enhanced deep feature representation for person re-identification. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8, March 2016.

[57] G. Xiao, W.-K. Fan, J.-F. Mao, Z.-B. Cheng, D.-H. Zhong, and Y. Li. Research of the fish tracking method with occlusion based on monocular stereo vision. In *2016 International Conference on Information System and Artificial Intelligence (ISAI)*. IEEE, June 2016.

[58] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[59] Z. XU and X. E. Cheng. Zebrafish tracking using convolutional neural networks. *Scientific Reports*, 7(1), Feb. 2017.

[60] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng. Patch-based discriminative feature learning for unsupervised person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[61] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3655–3661. AAAI Press, 2016.

[62] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[63] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *CoRR*, abs/1610.02984, 2016.

[64] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[65] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[66] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng. Deep hybrid similarity learning for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3183–3193, Nov 2018.